DOCUMENT RESUME

ED 408 332 TM 026 585

AUTHOR Fierros, Edward G.; And Others
TITLE Using Multiple Methods of Assessment To Promote District Level Reflection about Instructional Improvement..
PUB DATE 25 Mar 97
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Data Analysis; *Data Collection; Educational Change; *Educational Improvement; Educational Practices; Elementary Secondary Education; Evaluation Methods; Sampling; *School Districts; Student Attitudes; *Teachers; Urban Schools
IDENTIFIERS *Reflective Thinking; Reform Efforts

ABSTRACT

A common theme in current school reform efforts is that teachers within schools must become reflective practitioners if they are to become more successful in improving instruction to meet the needs of increasingly diverse populations. In an effort to help schools promote district-level reflection about instructional improvement, Boston College's Center for the Study of Testing, Evaluation, and Educational Policy (Massachusetts) assisted teachers in two urban districts to utilize an assessment approach that relied on multiple methods of gathering information about classroom practice. This approach suggests that schools seek alternative perspectives on the life of schools based on the insights and perspectives of those who are perhaps the most assiduous observers of school and classroom life, students. Survey responses from 1,402 students in one district and 720 in another were analyzed. The paper discusses the four fundamental components of the model: (1) involving practitioners in the design of assessments; (2) employing matrix sampling; (3) using multiple methods of assessment; and (4) involving practitioners in the interpretation of results. It provides examples of each of these key components from two districts. The paper also discusses the relative merits and limitations of using this model to promote district level reflection about instructional improvement. (Contains 21 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 408 332

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
Anne Wheelock
TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

USING MULTIPLE METHODS OF ASSESSMENT TO PROMOTE DISTRICT LEVEL REFLECTION ABOUT INSTRUCTIONAL IMPROVEMENT

by

Edward G. Fierros, Cengiz Gulek, Anne Wheelock
Boston College

This paper is presented at Annual Convention of the AERA; Chicago, Illinois, March 25, 1997.

Please address all correspondence to:

Anne Wheelock
Center for the Study of Testing, Evaluation and Educational Policy
323 Campion Hall
Boston College
Chestnut Hill, MA 02167

Tel. (617) 552-4521
Fax. (617) 552-8419
E-Mail: Wheelock@shore.net

BEST COPY AVAILABLE

TM026585

ABSTRACT

A common theme in current school reform efforts is that teachers within schools must become reflective practitioners if they are to become more successful in improving instruction to meet the needs of increasingly diverse populations (Schon, 1987, 1991; Sternberg & Horvath, 1995). In an effort to help schools promote district level reflection about instructional improvement, Boston College's Center for the Study of Testing, Evaluation, and Educational Policy assisted teachers in two urban districts to utilize an assessment approach that relied on multiple methods of gathering information about classroom practice. This approach suggests that schools seek alternative perspectives on the life of schools based on the insights and perspectives of those who are perhaps the most assiduous observers of school and classroom life, namely students. This paper discusses the four fundamental components of the model: (1) Involving practitioners in the design of assessments, (2) Employing matrix sampling, (3) Using multiple methods of assessment, and (4) Involving practitioners in the interpretation of results. It illustrates examples of each of these key components from two districts. Finally, this paper discusses the relative merits and limitations of using this model to promote district level reflection about instructional improvement.

USING MULTIPLE METHODS OF ASSESSMENT TO PROMOTE DISTRICT LEVEL REFLECTION ABOUT INSTRUCTIONAL IMPROVEMENT

INTRODUCTION

During the 1995-1996 school year, the Center for the Study of Testing, Evaluation, and Educational Policy (CSTEEP) at Boston College was involved in helping two urban school districts to assess the work of their middle schools, all of which were implementing the districts' standards-based reform strategy. The stimulus for this involvement was the districts' obligation to describe their work in middle school reform in an annual report for their funders and community constituencies. As part of our technical assistance related to the preparation of this annual report, we aided the districts in implementing a student survey to elicit middle grades students' views, attitudes, and experiences vis-à-vis standards-based reform. Prior to our involvement, the districts had gathered traditional quantitative data for assessing school progress, including scores on standardized tests, attendance rates, and dropout rates. Neither of the districts had sought to connect these data with classroom practices; neither had sought alternative perspectives into the life of schools and classrooms by drawing on the insights and perspectives of those who are perhaps the most assiduous observers of school and classroom life, namely students.

Several principles that had shaped CSTEEP's earlier work using such surveys guided our work with these districts. We believed the survey would be of utmost use when: (1) the assessment took place at the school level, (2) teachers could reflect on the patterns of student responses vis-à-vis their educational and instructional practices, and (3) teachers could interpret results in conjunction with other assessments such as performance, multiple-choice, and standardized assessments. Further, our approach to implementing the survey incorporated four key principles:

- Practitioners should be involved in the design of assessments;
- The survey should employ matrix sampling of students;
- The survey should use multiple methods to prevent any one medium from becoming the message;
- Results should be presented in an open-ended manner to practitioners in such a way as to involve them in interpreting results in the context of their own experiences.

In this paper, we describe the application of these principles in the two urban districts, which we call District A and District B. We also explicate the technical considerations in scoring a survey that includes both multiple-choice and open-ended questions and the ways in which we presented results to stimulate teacher reflection. Finally, we discuss the merits and limitations in applying this approach at the district level.

Involving Practitioners In The Design Of Student Surveys For School Assessments

We grounded our work with the two districts on a commitment to involve practitioners in designing the student survey. This commitment reflected the value we place on participatory planning and assessment as well as our understanding that practitioners are more likely to use information generated about schools if they have a hand in determining which information to collect (Patton, 1986). Given different conditions in the two districts and given that we were providing technical assistance at considerable distance, the form of involving practitioners varied between the two districts.

CSTEEP staff made on-site visits to each district three times over the course of the year. In both districts, initial visits involved meetings with key Central Office staff and an on-site review of available documents and data-management capacity.

Although these visits involved face-to-face consultation with district staff in both districts, specific staff involvement varied from one district to another. In District A, we met first with one Central Office staff member and her "Middle Level Advisory Group" consisting of one or two representatives from each middle school. In the initial meeting, we discussed the district's particular interests and presented examples from earlier survey results as a way of introducing the merits of using a three-part survey to assess student attitudes and classroom experiences related to middle school reform. Based on these discussions, the advisory committee selected questions from other national surveys, including the National Assessment of Educational Progress (NAEP) and New Standards, to gather information that could be compared to a national sample. In addition, because the district was especially interested in students' understanding of teachers' classroom standards, a second part of the survey incorporated a prompt that asked students to describe the differences they perceived between "excellent" and "very good" work. Subsequent discussions by mail, telephone, facsimile, and electronic mail between CSTEEP staff and the district's middle schools coordinator resulted in fine-tuning the survey and planning for administration.

In District B, CSTEEP staff first met with the district's academic coordinator and director of management information services to determine school-based data already available to schools. Following the initial visit, CSTEEP provided suggestions for "next steps" to Central Office staff, including that of forming a steering committee to plan for the annual report. On two subsequent visits, CSTEEP met with this steering committee, including Central Office representatives from academics, development, public relations, and data management, as well as one middle school principal and two middle school teachers. On each occasion, discussions involved the kind of data to be collected for the annual report, and during the first meeting of the group, we introduced the idea of gathering information about student perceptions of classroom practices to complement the school-based quantitative data that would be included in this report. After considering a variety of approaches to gathering students' perceptions about their learning experiences, we outlined a draft instrument. As in District A, fine-tuning of the survey and discussions about administration occurred in subsequent telephone conversations.

Ultimately, although we introduced the basic survey to practitioners in both sites, the surveys adopted and used were "custom-designed" to reflect the needs, interests, and political context particular to each district. In both districts, the final product reflected a negotiating process during which we introduced information about appropriate survey design and discussed the districts' willingness to break new ground. Risk-taking depended to some extent on the Central Office's willingness to answer concerns from stakeholders. For example, when union leadership in District B raised concerns about a student survey of classroom practices, district staff assured them that because teachers had participated in the survey design, teachers would accept the value of the survey. As a result of our discussions with educators, two of the three survey parts changed from the original. In both districts, educators selected some multiple-choice questions for the first part for reasons particular to the district. The second part of the survey also evolved from discussions in each district. In the third part, a prompt that encouraged students to draw one of their teachers at work in the classroom, was common to both districts.

EMPLOYING MATRIX SAMPLING OF STUDENTS

The idea of utilizing matrix sampling in this approach evolved from previous projects CSTEEP has assisted with. An extensive discussion of relative merits of this technique can be found in the report *Design for a New Generation of American Schools* by Bolt Beranek and Newman (1993). Without going into detail, the idea of incorporating matrix sampling suggests that in matrix sampling, the total pool of students is divided, and different but equivalent samples of students are surveyed. Thus, not all students are asked to complete the survey. Matrix sampling is

used to get accurate population estimates without having to survey each student. Matrix sampling is often used when there is not enough time or resources to administer to all students. In addition, using the matrix sampling procedure at the district level allows districts to generalize about schools without implicating specific teachers or children. Likewise, matrix sampling of open-ended questions generates richer data about classrooms that does not overwhelm the analysis. Moreover, by employing a sampling of grade levels, CSTEEP attempted to reduce the burden of external assessment on student.

Despite successful prior experiences using matrix sampling, differences in district conditions resulted in variations on this procedure. For example, District A altered the sampling strategy by administering the student reflection survey to all students. District A then drew random samples for each school regardless of the grade level in order to generalize results for each school as well as for the district as a whole.

District B, on the other hand, used a systematic random sample of four schools per grade level, with schools selected from different geographic areas of the city. District B selected random samples of grade levels and random samples of students within grades, providing a sample of 60 student surveys from each school. As the district's director of research and data management explained, this method of sampling was selected largely for convenience in response to year-end pressures and the timing of the survey, which was sandwiched in between statewide testing obligations. As in District A, this sampling method yielded a sample of 60 surveys from each school. Because of the sampling technique, the results could be generalized only at the district level.

USING MULTIPLE METHODS OF ASSESSMENT

In designing the survey, both District A and District B adopted the principle of using multiple modes of assessment to gather data from various perspectives and prevent any one assessment mode from determining "the message" of results. Thus, each district survey contained "Part A" multiple choice questions, requiring students to circle a response; a "Part B" with an open-ended prompt; and a "Part C" with a drawing prompt. However, again, the contents differed according to the different district contexts, with the drawing prompt being the only part that was consistent for each district.

For example, District A chose to use "Part A" of the survey to gather data from students that could be compared to national results. CSTEEP suggested sample questions form the National Assessment of Educational Progress (NAEP) student survey. In contrast, District B chose to use

"Part A" to determine students' perceptions of standards-based reform as it was evolving in that particular district. Curious about student responses to the district's recently adopted standards, the steering committee wrote five questions that asked students to describe the degree to which they agreed or disagreed with statements regarding the benefits of standards-based reform.

For the second part of the survey, District A asked students to respond to the following four statements:

1. Describe the things you like best about your school.
2. When your teachers read your essays and papers, how do they decide whether your work is "Excellent" (A) or "Very Good" (B)?
3. What are the most important things you have learned at your school this year?
4. Describe some suggestions you have for making your school even better.

In contrast, District B decided to use this part to provide teachers with information about those classroom activities that seemed most popular with students. The steering committee reviewed one approach that asked students to describe their most memorable learning experience (Wasserstein, 1995) and decided to write a similar prompt: "Describe the most memorable product/project you worked on this year." Finally, the prompt for "Part C" was common for both districts, asking student to "Think about the teachers and kinds of things you do in your classrooms. Draw a picture of one of your teachers working in his or her classroom."

Analyzing Data When Utilizing Multiple Methods of Assessment

Both district A and district B administered the student reflection survey late in the 1995-1996 school year, generating 1402 surveys from district A and 720 surveys from district B for analysis after sampling. Because neither district had the necessary time or staff available, both chose to have CSTEEP staff complete the data analysis and prepare results for presentation to practitioners. We conducted this analysis at Boston College through a series of both quantitative and qualitative data analysis.

For Part A, we performed straightforward quantitative analyses to summarize percentages of students choosing each response category. This part was machine scorable, with analysis yielding a picture of student agreement with certain attitudinal items reported in percentages. In District A, the results were compared to national data. For example, the district used a nationally normed question on a sample of 1,402 middle schoolers. The question was: "There is a good communication between students and teachers." We computed the descriptive statistics on this

question for the district. The maximum, minimum, median and average, and standard deviation scores were 70%, 36%, 59%, 57%, and 10%, respectively. The national average was 42%. Thus, District A can claim relatively "better communication" between students and teachers, about 15% higher than the national data. We also reported Part A survey results for District B as percentages, but since the questions were particular to the district, we could compare results for different grade levels districtwide only.

In Parts B and C, we systematically reviewed student responses to identify general patterns as well as specific characteristics. Multiple independent raters, provided with standardized guidelines for coding and scoring, carried out qualitative data analysis to systematically review student responses both holistically and analytically. The holistic method of scoring entailed awarding a single score to each general pattern based on the overall impression, whereas the analytic method broke down the general patterns into subcategories, each of which is scored independently (Mills, 1991; Airasian, 1995; Linn and Gronlund, 1995). We examined holistically general patterns such as whether specific technology, people, or physical features were contained in a school district. We scored the specific characteristics such as computers, math class, teachers, or cooperation from an analytical perspective.

Reliability Considerations in Utilizing Multiple Methods of Assessment

In assessments where free responses are scored according to criteria, it is essential to have consistency, better known as reliability, among those who score the responses (Airasian, 1994; Linn & Gronlund, 1995). Two sections in the student reflection survey, the open-ended and drawing sections required subjective judgments to score student responses and used multiple raters to score student responses. We examined the consistency among raters (inter-rater reliability) as well as the consistency within one rater (intra-rater reliability) to ensure reliability.

In scoring occasions of nominal data where there is a substantial proportional gap between two categories, Cohen's Kappa adjustment for the coefficient of agreement is considered the most appropriate fit for the situation (Burton, 1981). Feingold (1992) indicates that Cohen proposed Kappa in order to adjust gross agreement by considering the extent of agreement that would occur by chance, because of each judge's overall, or marginal, assignments to each category of the rating scale. Chance agreement, as defined by Feingold, refers to the proportion of times that two judges (or raters) would be expected to agree if their ratings were independent of each other. An estimate of Kappa, using sample proportions would be the ratio of the difference of proportion between the observed and the expected agreement to the subtraction of the proportion of expected agreement from 1 (Kvalseth, 1989; Feingold, 1992).

We show an example of our extensive inter-rater reliability analysis for the open-ended and drawing sections of the survey in Table 1. To analyze the data collected from District A in Spring 1996, three independent raters scored thirty-seven randomly selected student surveys. We then cross-checked the ratings in pairs. A coefficient of agreement as well as the adjusted coefficient of agreement (i.e., Kappa) were reported for each open-ended and drawing item in Table 1 below.

Table 1. Inter-Rater Reliability Coefficients of Agreement.

	Rater 1	Rater 2	Rater 3
Open-Ended Item 1			
Rater 1	1.00		
Rater 2	.85 [1] [.99 [2], .91 [3]]	1.00	
Rater 3	.95 [.99, .92]	.81 [.98, .91]	1.00
Open-Ended Item 2			
Rater 1	1.00		
Rater 2	.85 [.99, .93]	1.00	
Rater 3	.91 [.99, .92]	.86 [.99, .92]	1.00
Open-Ended Item 3			
Rater 1	1.00		
Rater 2	.80 [.99, .93]	1.00	
Rater 3	.82 [.99, .93]	.86 [.99, .93]	1.00
Open-Ended Item 4			
Rater 1	1.00		
Rater 2	.76 [.98, .93]	1.00	
Rater 3	.85 [.99, .93]	.81 [.99, .97]	1.00
Drawing Item 1			
Rater 1	1.00		
Rater 2	.85 [.96, .71]	1.00	
Rater 3	.81 [.94, .71]	.82 [.95, .72]	1.00

Note. [1]: Cohen's Kappa Coefficient of Agreement (Adjusted).
[2]: Observed (Overall) Agreement.
[3]: Expected Agreement.

Note that results show a substantial change for the coefficients of agreements before (simple percent agreement) and after the Kappa adjustments were made. For instance, Open-ended item 1, shown in Table 1, has an observed coefficient of agreement between rater 1 and rater 2 that is quite similar observed coefficient of agreements between rater 1 and rater 3. Thus, on the surface, there seems to be no difference among the three raters in terms of the observed and the expected agreement. However, the Kappa coefficients of agreement showed that rater 1 has higher level of agreement with rater 3 than rater 2 (about 10% difference), when the agreement among raters is adjusted for the chance factor.

Table 1 shows that in general, the coefficients of agreement were quite high for all questions in the reflection survey. According to Kvalseth (1989), Kappa coefficient of .61 is a reasonably good over-all agreement. The lowest and the highest Kappa coefficients in the reflection form are .76 (question 4, between rater 1 and rater 2) and .95 (question 1, between rater 1 and rater 3), respectively (see also Table 1). Thus, we were able to attain a high degree of consistency in scoring.

Because the student reflection survey may be scored at different times for a school district and/or same raters may be scoring surveys from different school districts administered in different periods of time, it is important to investigate how the scoring within one rater changes over time. Measuring the consistency of scoring within a rater is also possible with the intra-rater reliability technique. The process requires selecting a sample of student surveys and scoring them by the same rater two or more times with a certain amount of time in between.

We studied the consistency within one rater by taking 61 randomly selected student surveys and asking the same rater to score them twice, with a time interval of two weeks between two scorings. The observed coefficient of agreement and the adjusted Kappa coefficient of agreement were .97 and .89, respectively, providing a highly satisfactory intra-rater reliability coefficient.

Validity Considerations in Utilizing Multiple Methods of Assessment

The essence of content consideration in validation, as explained by Hopkins, Stanley and Hopkins (1990), and Linn and Gronlund (1995), is determining the adequacy of sampling of the content that the assessment results are interpreted to represent. The goal in the consideration of content validation is to determine the extent to which a set of assessment tasks provides a relevant and representative sample of the domain of tasks about which interpretations of assessment results are made (Linn and Gronlund, 1995).

The content considerations in validating the reflection survey involved negotiating with schools over the kind of information teachers thought would best provide a profile of the school and stimulate teacher reflection on their instructional practices. For example, in line with district reform initiatives, District A teachers were interested in knowing the extent to which students were aware of standards, and how they understood the difference between an "excellent" and "very good" work. As a result, District A adapted one open-ended question prompt: "When your teachers read your essays and papers, how do they decide whether your work is excellent (A) or very good (B)?" In District B, teachers were interested in the kinds of classroom assignments that most engaged their students. Thus, they adapted a question from an article that described how one teacher had addressed this concern in another district.

The scoring of the Student Reflection Survey requires that open-ended and drawing questions be coded by trained raters. The student responses are coded in terms of certain characteristics they to arrive at a general pattern/category of responses. Some examples of general categories were Technology, Subjects, People, and Activities; and some examples for individual characteristics would be Internet (under Technology), Writing (under Subjects), Athletics (under Activities), and Principal (under People). Many of these characteristics have commonality across districts since every school has subjects (Math, Science, Reading, and so on) in the coding/scoring. All of these general characteristics make up the construct that is being investigated.

We addressed the construct validity of the survey approach by using expert judgments in the definition of constructs. Five experts Master's and doctoral students in Educational Research, Measurement and Evaluation (ERME) program at Boston College were provided with a sheet containing randomly ordered specific characteristics and a list of general categories. To illustrate, one of the general categories was "technology" which has 7 characteristics (or sub-categories) such as Computers, TV/VCR, Internet/WWW, E-Mail, Software Applications, Software Titles, and Technology in General. The task of the experts was to assign each characteristic to a given general category according to the descriptions of individual characteristics and general categories which were provided to experts as a reference during categorization. Overall, there were 50 individual characteristics and 9 general categories to be matched by the experts. The percentage of minimum and maximum correct matchings were 72 and 90, respectively; with a percent mean correct matching of 82. Thus, on the average, raters correctly identified 82% of the individual characteristics to belong to a general construct.

Just as assessments are intended to contribute student learning, student survey results are intended to affect classroom practice. In this vein, Messick (1989) suggests that the overall judgment of validity of particular uses and interpretations of assessment results requires an

evaluation of the consequences of those uses and interpretations. The intended use of student survey results in Districts A and B was to stimulate teacher discussion of the patterns teachers perceived in the open-ended and drawing responses of the survey and identify possible instructional improvements suggested by those patterns. In an effort to determine if the survey would result in teachers' describing patterns that could result in more reflective practice, we intentionally delivered same examples of two small groups of teachers who analyzed the same set of drawings that CSTEEP raters had scored. The groups came up with the following conclusions:

	Group Conclusions	**CSTEEP Findings**
Group 1: **School X**	• Traditional Classroom Settings • Teacher at Front of the Blackboard • No Evidence of Technology • Students at Desks • If Talking Represented, It's the Teacher • Almost all Positive Depictions, With Two Exceptions • Board Assignments Not Innovative or of Substance	• Teacher Depicted at the Blackboard (42%). • Teachers at Teacher Desk (33%) • Teacher Depicted Alone (57%) • Student at Desks (35%) • Teacher Desk Depicted(47%) • Student Desks in Rows (23%) • Teacher Positive Demeanor (47%) • Teacher Negative Demeanor (8%). • Computers Depicted (3%)
Group 2: **School X**	• Whole Classroom Drawings • Lots of Examples of Politeness • Smiling Faces • Technology Depicted More Than Few Times • Some Negatives, Lots of Positives • Different Seating Patterns	

As represented in the table above, although the two teacher groups observed a variety of patterns, both groups indicated that the collection of drawings had many positive features (such as smiley faces) and little or no evidence of technology. The analytic scoring of drawings show similar patters to those indicated by teachers in small groups. For example, both groups identified "traditional classroom setting," which is described as the teacher standing at the blackboard alone, or sitting at his/her desk and student desks are in rows. Indeed, about 42% of students at this school depicted teachers at the blackboard, 33% depicted teachers at the teacher desk, and 57% of drawings had teachers alone in the picture. Also, 35% of students were depicted at their desks. In classroom setting, 47% of student drawings included teacher desk, and 23% included student desks in rows. Two common observations by teachers were positive classroom atmosphere, yet little or no evidence of technology were also backed up by Boston College's assessment team:

about 47% of student drawings had teacher as positive demeanor, whereas only 8% had negatives; only about 3% of the drawings included computers. The substantial degree of correspondence between teachers' holistic interpretations and the analytic interpretations by Boston College the potential for teachers to interpret results in ways that could result in improved classroom practice.

PRESENTING ASSESSMENT RESULTS TO PROMOTE REFLECTION ON CLASSROOM PRACTICE

In both District A and District B, the student reflection survey, especially the open-ended and drawing prompts, served the purpose of providing an entry point into discussions with educators about teaching and learning experiences in their own school's classrooms. We initiated these discussions in half-day workshops convened by Central Office staff. These workshops allowed us to model our data analysis process for district educators before distributing the results as analyzed by CSTEEP staff to each school. In both districts, the large group attending the workshop included both principals and teachers, with several district staff also attending. In District A, the meeting was voluntary but drew participants representing each school; in District B, Central Office required the principal and at least one teacher required to be present. In these workshops, we provided attendees with representative responses from the district along with the scoring rubrics used to record data for Parts B and C. Using these responses, we asked educators to work in pairs or school groups to review and score responses.

After allowing principals and teachers time to review the surveys and begin recording responses, we asked each school team to reflect on the patterns they were observing, the reasons these patterns might occur, and the kinds of things they might do differently as a result. We solicited responses from each team for discussion with the entire group. We again made explicit our hope that principals would replicate this process with teachers in their own schools. Only after we had walked participants through the reviewing and scoring process did we distribute our own analysis of data from each school.

At the end of each workshop, we also solicited written feedback from those attending. In addition, we asked each principal to provide us with responses from their faculty after they had used the surveys in their own schools according to the process we had modeled.

The responses we received suggested both strengths and weakness of this tool to promote teacher reflection on classroom practices. On one hand, responses indicated some openness to rethinking classroom practices to address concerns raised by student responses. For example, in response to the prompt: "Describe the most memorable product/project you worked on this year,"

teachers indicated greater awareness of students' positive reaction to groupwork. Some connected this with social needs of young adolescents: *"What I noticed is that students find projects that involve group work the most memorable. I think this has a lot to do with the age group of the students surveyed,"* said one; "*In middle school, students' interaction with peers is a priority*," said another. Others observed, "*Students working in interdisciplinary groups seemed more excited than a traditional class atmosphere*" and, simply, *"[I noticed that] students prefer working in groups."*

Teachers' examination of responses to this question also pointed to subject areas and assignments that had most engaged students. Thus, one principal reported, *"Reading and writing projects are viewed in a positive manner by students [in our school]."* In other schools, responses seemed to raise educators' awareness of the value students placed on hands-on and project learning, as in the following comments:

* *"The matrix [coding sheet] gave us insight to various activities that are common in classrooms such as: product based activities, team-work, hands-on activities and research. It helps the teachers gain a better understanding of what techniques are memorable to students."*

* *"Projects and presentations are seen as positive, motivational, and memorable by students!!"*

* *"In the open-ended response about 'memorable' product/process students again indicate direct involvement is important to them. Their most 'memorable' work was work which stretched over time and/or involved them with other people (students), or involved them in hands on activities."*

* *"Students like activities which allow them to have hands on experience."*

* *"The extend[ed] projects appear to be more meaningful than those which were short term."*

* *"The projects provide greater ownership of their learning and are different from their past experiences in learning."*

On the other hand, although one educator noted, *"Student perception of teachers is very enlightening,"* principals' and teachers' reactions also indicated less willingness to entertain less flattering student comments. In particular, student drawings that negatively portrayed classrooms as being dominated by "teacher talk" or hostile student-teacher interaction elicited alternative

explanations for student responses. For example, several principals took the abundance of student drawings depicting teachers alone at the blackboard as signs that students were adept at following directions quite literally. As one principal noted, "*The patterns occur because of the way the questions are asked. Our state test's writing instructions have caused students to be focused on their cues, and they are pretty good at it.*" Another explained:

> *"If the drawing had not been specifically requiring a teacher, it would have contained more details about students themselves. Since the requirement was stated the way it was, students depicted school as teacher centered and teacher dominated. However, the students knew how to follow directions and drew the teacher dominating the scene."*

One teacher put drawings that negatively portrayed teacher behavior in the context of early adolescence, noting:

> *"Students see this as an opportunity to 'cartoon' and therefore the drawings do not provide any substantive information."*

Others elaborated on specific classroom conditions that prevented them from abandoning traditional practices. For example, one educator noted, *"More group work is needed. [But] projects completed in class are difficult to accommodate because there are no funds for materials (especially when our class numbers over 30)."* And another reported:

> *"On Part C, I noticed that several students drew pictures of teachers addressing the whole class. I think this has [to do] with the phrasing of the questions and that with large classes, teachers do have to spend some their time using whole group instruction."*

Some speculated that student drawings were less a reflection of students' experiences in their current schools than the result of experiences accumulated over six to eight years of prior schooling. One principal brought the school's survey results back to his faculty and reported:

> *"The council believed that high number of responses of teacher depicted alone, teacher drawn as full-figure, and at the blackboard or at desk and students desks in rows seems to follow the typical stereotype and conditioning of the student for the 'normal' classroom (drawn from the students' last seven years of schooling)."*

Another added:

> *"The survey indicated that the traditional seating patterns and teaching methods are dominant. Whereas we agree that the method of delivering instruction is probably very close to accurate, there traditional seating arrangements of the rooms, visible throughout the building is not reflected."*

In our workshop, we had little time to engage teachers in extended discussion of these responses. However, we did have time to validate these as legitimate reactions and to note that in other schools, when teachers had used the same survey over several years, the student responses changed as classroom practices changed, even though the question was worded exactly the same (Haney, Russell, & Sack, 1996).

At the same time, some educators indicated in their feedback to us that they would take the results of the student surveys into account in future planning. For example, several principals indicated that they would focus attention on student-teacher interaction as in the following comments:

> * *"We as a campus need to spend time reflecting on what we are portraying as important to students. Products could be re-designed and teacher demeanor needs to be addressed."*
>
> * *"Over half of the staff appears enthusiastic in their teaching (53%), whereas 47% appear unhappy, or no emotion. (This finding will be addressed by the campus administrators.)"*

Other principals suggested they would attempt to help teachers make specific structural changes in classrooms. One noted:

> *"Based on the responses on the survey teachers may want to allow the students more freedom to move around the classroom as a learning tool. Teachers may want to allow for more students directed learning experiences and the teacher used as a facilitator. The teacher may wish to develop a less rigid teaching style, more group activity, and activities to allow for the different learning styles."*

Others offered a list of specific steps they intended to take to respond to student comments and drawings, including:

> *"Greater emphasis on classroom environment in relation to seating arrangement. Even greater emphasis on projects. The school will support these changes through various methods:*

A. teacher training in teaching strategies
B. staff meetings organized using the methods desired for the teachers to use
C. greater emphasis on cooperative and peer teaching (as well as other strategies related to learning styles)
D. staff development funds dedicated to the support of the support of the areas in need
E. continuance of the emphasis on the Academic Standards, especially those visible to the students
F. provide the teachers more time to develop the projects for each standard."

And:

" Arrange desks in groups/clusters*
** Teacher needs to move around the room.*
** Projects/Products need to be done in all subject areas.*
** Bulletin Boards need to be meaningful to students.*
** Computers and other media materials need to be used daily.*
** Students need to work in groups."*

Not all principals took the student survey results as a call for change. In fact, the question, "What will your school do differently as a result of this survey?" elicited ambivalence about what changes could or should be made. As one principal reported:

"We are trying to relate education to the 'real world' with our emphasis on performance standards. We are trying to provide ways for students to interact with each other, become aware of how they learn, and take responsibility for their own advancement. I am not sure we need to do things 'differently.' We need time to do the things we start before being asked to do new things."

DISCUSSION

In recent years, literature on education reform has emphasized that schools must become places where teachers can engage in critical study about their own practices (Darling-Hammond, 1988; Glickman, 1993; Sirotnik, 1987; Sirotnik and Oakes, 1990). Our experience suggests that districts can use Student Reflection Surveys as one tool to assist educators at the school level in assessing their own practice. At a time when school accountability policies emphasize student outcome data, this survey can add balance to a picture of school and district practice by providing student perceptions of teaching and learning.

In fact, in sessions where we worked with educators to interpret results, we were impressed with how powerfully educators reacted to the open-ended responses in particular. Perhaps the visual data of student drawings has the capacity to penetrate teachers' consciousness in a way that numerical data on its own cannot do.

By focusing on student attitudes and experiences, survey results can also delineate the characteristics of the classroom context that may be affecting student performance data at the school level. Further, at the district level, results can assist district leadership in reassessing policy initiatives, allocating resources, and designing professional development opportunities. Results could also alert district staff to strengths and weaknesses of particular schools, and the process of interpreting data can offer an opportunity for district and school personnel to work together to rethink teaching strategies. Indeed, our experiences working with these districts revealed how infrequently district-and school-level educators meet to discuss classroom practice. The convening of staff from the two levels to discuss survey results was unique in this regard.

However, although we believe the survey has promise, our experience also suggests that effective use of this tool is contingent on other conditions. Our assistance to the two districts took place over a period of less than a year, during which we developed a relationship with district staff, provided technical data analysis, and facilitated interpretation of results with school-level practitioners. The timing of our project did not allow more follow-up with individual schools. Thus, we do not know how the results have been used at the school level.

At this point, it is premature to predict the extent to which the districts will support on-going use of the survey. On one hand, the districts have incorporated results of the survey, including summaries of multiple choice results, summaries of responses to open-ended questions, and sample drawings, in their annual reports to funders, business leadership, and other community constituencies. On the other hand, without technical assistance or pressure from funding sources, the districts may see the survey as an interesting experiment, but one that requires more resources than they have. For example, the data management and research departments in both districts are thinly staffed and already burdened by reporting requirements and state-level accountability data. Likewise, both district-level and school-level reflection requires time for educators to meet together to discuss survey results, and incentives to use extra time for reflection purposes do not exist.

An additional barrier to using the survey at the district level resides in the political realm. Over the course of our work with the districts, we became aware of the intense pressures on the districts to produce "good news" about middle school reform for public relations purposes. This pressure in a context of high-stakes accountability policies leaves little incentive for districts to promote a process of data-gathering that threatens to reveal fundamental problems students face in classrooms. Districts' need to put the most positive face on reform can affect the design of the survey. Moreover, if district staff are preoccupied with school accountability, they are unlikely to see the survey as a useful tool.

In the context of pressures to look good, districts may be less prepared to entertain data that suggest negative student experiences. And given the defensiveness that arises when teachers become aware of students' concerns, schools may make use of the survey data for critical inquiry only if the districts provide leadership by making the student survey an annual event and providing direction for reflection among faculty in a safe, no-stakes context. While we believe that multi-year use of the survey could result in school-based staff seeing changes in student attitudes and experiences as reforms take hold, we can not now predict that district leadership will make the survey process part of its standard operating procedure and allocate Central Office resources to support data gathering and facilitate school-based inquiry in future years.

The student reflection surveys that evolved from our work remain, in our view, only one form of gathering information about the work of schools. Its unique contribution is that it is a multi-faceted vehicle for focusing on student attitudes and experiences related to classroom practices. Districts could tap its full potential by combining it with a more comprehensive assessment program, with an emphasis on the survey as a tool for critical inquiry into instructional improvement rather than for external evaluation and/or school accountability.

REFERENCES

Airasian, P. W. (1994). *Classroom Assessment*.. New York: McGraw-Hill.

Bolt, Branek, Newman Inc. (1993). *Design for a New Generation of American Schools: Milestone 2 Deliverables*. Report Submitted to The New American Schools Development Corporation.

Burton, N. W. (1981). Estimating Scorer Agreement for Nominal Categorization Systems. *Educational and Psychological Measurement*, *41*, 953-962.

Darling-Hammond, L. (Winter 1988). Accountability and Teacher Professionalism. *American Educator*, *12*, 8-13.

Feingold, M. (1992). The Equivalence of Cohen's Kappa and Pearson's Chi-Square Statistics in the 2X2 Table. *Educational and Psychological Measurement*, *52*, 57-61.

Finn, C. (1991). *We Must Take Charge: Schools and Our Future*. New York: Free Press.

Glickman, C. (1993). *Renewing America's Schools: A Guide for School-Based Action*. San Francisco: Jossey Bass.

Haney, W. & Raczek, A. (1994). *Surmounting Outcomes Accountability in Education*. Report Prepared for the U.S. Congress Office of Technology Assessment.

Haney, W., Russell, M. & Sack, T. (1996). *Reflecting Educational Practice: Using Student Drawings to Illuminate the Educational Ecology of Schools and to Promote Teacher Reflection*. Research Proposal to the Spencer Foundation from the Center for the Study of Testing, Evaluation and Educational Policy at Boston College.

Hopkins, K. D., Stanley, J. C. & Hopkins, B. R. (1990). *Educational and Psychological Measurement and Evaluation*. New Jersey: Prentice Hall.

Kvalseth, T. O. (1989). Note on Cohen's Kappa. *Psychological Reports*, *65*, 223-226.

Linn, R. L., Gronlund, N. E. (1995). *Measurement and Assessment in Teaching*. New Jersey: Prentice-Hall, Inc.

Patton, M. Q. (1986). *Utilization-Focused Evaluation (2nd Ed.)*. Beverly Hills: Sage Publications.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement (3rd Ed.)*. New York: American Council on Education/Macmillan.

Mills, J. (1991). Assessing Musical Performance Musically. *Educational Studies*, *17* (2), 173-181.

Schon, D. (1987). *Educating the Reflective Practitioner*. San Francisco: Jossey-Bass.

Schon, D. (1991). *The Reflective Turn: Case Studies in and on Educational Practice*. New York: Teachers College Press.

Sirotnik, K. A. (1987). Evaluation in the Ecology of Schooling: The Process of School Renewal. In J. I. Goodlad (ed.), *The Ecology of School Renewal: 86th Yearbook of the National Society of the Study of Education*. Chicago: University of Chicago Press.

Sirotnik, K. A. and Oakes, J. (1990). Evaluation as Critical Inquiry: School Improvement as a Case in Point. *New Directions for Program Evaluation*, *45*, 37-60.

Sternberg, R. J. & Horvath, J. A. (1995). A Prototype View of Expert Teaching. *Educational Researcher*, *24* (6), 9-17.

Wasserstein, P. (1995). What Middle Schoolers Say About Their Schoolwork? *Educational Leadership*, *53*, 41-43.